Bioinformatics, 36(26), 2020, i592–i600
doi: 10.1093/bioinformatics/btaa798
ECCB2020

Genes

An efficient framework to identify key miRNA–mRNA regulatory modules in cancer

Milad Mokhtaridoost[1] and Mehmet Gönen[2,3,4,*]

[1]Graduate School of Science and Engineering, İstanbul 34450, Turkey, [2]Department of Industrial Engineering, College of Engineering, İstanbul 34450, Turkey, [3]School of Medicine, Koç University, İstanbul 34450, Turkey and [4]Department of Biomedical Engineering, School of Medicine, Oregon Health & Science University, Portland, OR 97239, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Micro-RNAs (miRNAs) are known as the important components of RNA silencing and post-transcriptional gene regulation, and they interact with messenger RNAs (mRNAs) either by degradation or by translational repression. miRNA alterations have a significant impact on the formation and progression of human cancers. Accordingly, it is important to establish computational methods with high predictive performance to identify cancer-specific miRNA–mRNA regulatory modules.

Results: We presented a two-step framework to model miRNA–mRNA relationships and identify cancer-specific modules between miRNAs and mRNAs from their matched expression profiles of more than 9000 primary tumors. We first estimated the regulatory matrix between miRNA and mRNA expression profiles by solving multiple linear programming problems. We then formulated a unified regularized factor regression (RFR) model that simultaneously estimates the effective number of modules (i.e. latent factors) and extracts modules by decomposing regulatory matrix into two low-rank matrices. Our RFR model groups correlated miRNAs together and correlated mRNAs together, and also controls sparsity levels of both matrices. These attributes lead to interpretable results with high predictive performance. We applied our method on a very comprehensive data collection by including 32 TCGA cancer types. To find the biological relevance of our approach, we performed functional gene set enrichment and survival analyses. A large portion of the identified modules are significantly enriched in Hallmark, PID and KEGG pathways/gene sets. To validate the identified modules, we also performed literature validation as well as validation using experimentally supported miRTarBase database.

Availability and implementation: Our implementation of proposed two-step RFR algorithm in R is available at https://github.com/MiladMokhtaridoost/2sRFR together with the scripts that replicate the reported experiments.

Contact: mehmetgonen@ku.edu.tr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancer is one of the most leading causes of death worldwide. Although significant improvements have been made in cancer therapies in recent years, the molecular mechanisms of cancer have not yet been fully understood (Zhang *et al.*, 2019).

Messenger RNAs (mRNAs) are an immense family of RNA molecules that carry genetic information from DNA to the ribosome, where they specify the amino acid sequence of the protein products of gene expression. Micro-RNAs (miRNAs) are non-protein-coding RNA molecules found in plants, animals and some viruses that function in mRNA silencing and post-transcriptional regulation of gene expression through translational repression or mRNA degradation (Bartel, 2004). Considerable advances in the detection of miRNA characterization such as those conducted by He and Hannon (2004) and Raad *et al.* (2020) showed that these small non-coding RNAs are an important family of regulatory RNAs. It is a widely held view that miRNAs involved in the most biological processes, including the regulation of pathways that are effective in the formation and progression of many cancer types (Calin and Croce, 2006). Despite the significant importance of the miRNAs and their combinatorial effects as a class of regulatory molecules, their particular functional roles in cellular processes are still uncertain. Mentioned evidences clearly indicate a need to understand the regulation of mRNAs by miRNAs in cancer cells.

1.1 Previous studies

Even though the complex interactions between miRNAs and mRNAs are unknown, miRNA and mRNA expression profiles from the same sample provide a great opportunity to identify miRNA–mRNA interactions. In this context, several computational techniques have been proposed by applying probabilistic approaches,

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/36/Supplement_2/i592/6055910 by guest on 19 February 2021

machine learning techniques and statistical models to find miRNA–mRNA interactions using their expression profiles.

A large amount of literature has been published to explore cancer-related miRNA–mRNA interactions by applying probabilistic methods (Bonnet *et al.*, 2010; Joung *et al.*, 2007; Li *et al.*, 2014). This kind of probabilistic methods identify miRNA–mRNA interactions with a limited number of features, but they provide poor interpretation and limited information due to the high-dimensional and correlated input data.

To extract more accurate information about miRNA functions in gene regulation, several statistical data mining methods have been developed in recent years to learn the complex relationships between miRNAs and mRNAs (i.e. miRNA–mRNA regulatory modules).

Several studies have been focused on non-linear matrix factorization (NMF) methods to develop statistical data mining models for identifying miRNA–mRNA regulatory modules from their expression profiles (Shi *et al.*, 2017; Xiao *et al.*, 2019; Zhang *et al.*, 2011). These NMF-based algorithms conjointly estimate low-rank matrices for each source of input data (i.e. miRNA expression profile, mRNA expression profile, etc.) and then select co-modules from low-rank matrices using different strategies such as score- and threshold-based approaches.

A large collection of studies that applied statistical models for identifying miRNA–mRNA regulatory problems have emphasized the use of methods based on singular value decomposition (SVD) to model high-dimensional complex relationships of miRNAs and mRNAs in multivariate regression models (Koo *et al.*, 2018; Ma *et al.*, 2014; Min *et al.*, 2016; Wang *et al.*, 2015). Multivariate regression analysis is a technique that focuses on modeling the relationships between several real-valued outcome variables, known as *responses* (i.e. mRNAs) and a set of real-valued input vectors, known as *predictors* (i.e. miRNAs) or features. Ma *et al.* (2014) formulated a factorization model named T-SVD, based on SVD with a thresholding step. They applied a Bayesian information criterion-based approach for model selection. Wang *et al.* (2015) proposed a regularized regression based on adaptive Lasso to find miRNA–mRNA interactions. Min *et al.* (2016) proposed a two-stage method that first builds the regulatory network between miRNAs and mRNAs and then uncovers regulatory modules from the learned network using an ℓ_0-penalized SVD model. Koo *et al.* (2018) formulated a multivariate regression model using a method based on artificial neural networks to find miRNAs and mRNAs involved in regulatory modules and predicted their weights in miRNA–mRNA regulatory modules in two separate steps.

Here, we note that the interpretability of results and the predictive performance of all the aforementioned existing statistical data mining methods, which are proposed for identifying miRNA−mRNA regulatory modules, are examined on a limited number of cancer cohorts.

In many high-dimensional biology applications such as studies involving gene expression data,where the number of predictors (i.e. miRNAs) and/or responses (i.e. mRNAs) are considerably greater than the sample size, most of the standard procedures such as linear regression cannot consistently estimate a general high-dimensional vector of variables, thus the additional side-information is needed to impose low-dimensional structure on the model (Peng *et al.*, 2010).

To fit models on such datasets, regularized or penalized methods are needed, and variable selection/feature extraction is often the most important aspect of the analysis. Applying ℓ_1-regularized methods and linear factor regression are two fundamental approaches to estimate the parameters and impose sparsity structure on high-dimensional low sample size genomic data (Ajana *et al.*, 2019; Hao *et al.*, 2017).

Feature selection can be performed using ℓ_1-regularized methods such as least angle selection and shrinkage operator (Lasso). These methods were proved to be efficient in several applications that exhibit some types of sparsity. Unlike the classical variable selection methods such as subset selection, ℓ_1-regularized methods are computationally feasible for high-dimensional data (Zhang and Huang, 2008). In such regression applications (e.g. large-scale genomic studies), it is quite applicable and beneficial to impose low-dimensional structure on the coefficient matrix to decrease the adverse effect of the high dimensionality.

In regression problems, where an underlying factor contributes multiple predictors that are distinct yet related, the coefficient indices generally appear in groups. Grouping can be observed in several genetic applications, for instance, in miRNA–mRNA regulation, miRNAs (mRNAs) that are correlated to each other could be reported together. Selecting (or setting to zero) a whole group together as a unit by applying mixed ℓ_1/ℓ_γ norm ($\gamma > 1$) in a regularized method can capture the structured sparsity of the problem (Breheny, 2015).

Instead of approaches that impose constraints on the regression coefficient entries directly, another common strategy to induce sparsity is reducing the dimensionality by forcing a constraint on the rank of the coefficient matrix, which leads to reducing the number of variables. These methods are known as linear factor regression. A variety of well-known approaches including principal components regression (Massy, 1965) and SVD (Golub and Reinsch, 1971) can be included in linear factor regression models.

Various linear regression methods have been introduced to impose dimensionality reduction by writing the coefficient matrix as the multiplication of two smaller matrices, based on projecting the input data matrix into a low-dimensional space using the first matrix and performing linear regression in this projected space using the second matrix. The ranks of these matrices are smaller than or equal to the dimensionality of the projected space (i.e. latent factors), thus the rank of the coefficient matrix cannot be greater than the number of the hidden factors. Applying this approach can bring a major advantage by substantially reducing the number of variables, leading to a more sparse structure. Several studies have applied linear factor regression in a variety of ways. For instance, Argyriou *et al.* (2008) imposed an orthogonality constraint on the first matrix and placed $\ell_{2,1}$ penalty for the second matrix. Kumar and Daume (2012) performed ℓ_1 penalty on the first matrix and Frobenius penalty on the second matrix as the regularization term. Chen and Huang (2012) proposed a sparse reduced-rank regression method where $\ell_{1,2}$ penalty is applied on the first matrix and an orthogonality constraint is imposed on the second matrix. Kharratzadeh and Coates (2016) proposed a sparse multivariate factor regression method that enforces sparsity with an elastic net penalty on the first matrix and an ℓ_1 penalty on the second matrix.

In this paper, we developed an efficient two-step framework to find cancer-specific miRNA–mRNA regulatory modules from matched miRNA and mRNA expression profiles of tumors. In the first step, we estimated the interaction matrix of miRNAs and mRNAs using their expression profiles. In the second step, we designed a unified regularized factor regression (RFR) model that decomposes interaction matrix into two low-rank matrices with a separate elastic net penalty term on each. The first matrix linearly transforms the predictors (i.e. miRNAs) to a set of latent factors, and the second one regresses the responses (i.e. mRNAs) using these factors. Our model simultaneously estimates the number of effective latent factors (i.e. miRNA–mRNA modules) as a variable, applies low-rank structure on miRNA–mRNA interaction matrix and extracts miRNA–mRNA regulatory modules. Moreover, RFR model controls the sparsity level of both low-rank matrices separately and groups correlated miRNAs together and correlated mRNAs together. These attributes lead to interpretable results with high predictive performance.

1.2 Our contributions

To the best of our knowledge, the aforementioned linear factor regression models that proposed to identify miRNA regulatory modules (e.g. miRNA–mRNA, miRNA–disease, miRNA–lncRNA, etc.) considered the number of modules (i.e. the number of latent factors) as an input parameter and selected it using cross-validation (in the presence of cross-validation error) or rank selection approaches. These methods select the number of latent factors based on the input data (not the regression model), while regularization part of the model that includes regularization parameter(s) can alter the number of effective latent factors. Note that large and small numbers of

Downloaded from https://academic.oup.com/bioinformatics/article/36/Supplement_2/i592/6055910 by guest on 19 February 2021

latent factors would lead to overfitting and underfitting, respectively. To address this issue and improve the predictive performance of our proposed approach, we estimated the effective number of latent factors by incorporating it in the RFR model.

To extract the identified regulatory modules with biological relevance (i.e. key regulatory modules), we filtered the extracted regulatory modules by performing functional gene set enrichment analysis and survival analysis. To validate the regulatory modules identified by our algorithm, we performed literature-based validation and validation using experimentally supported miRTarBase dataset (Chou *et al.*, 2018). We tested our computational approach on a very large collection of diseases by performing computational experiments on 32 cancer cohorts. Our algorithm outperformed T-SVD regression in 29 out of 32 cohorts in terms of normalized root mean squared error (NRMSE) and extracted interactions with higher validated percentage in 17 out of 32 cohorts. The results showed that (i) identified modules are significantly enriched in `Hallmark, PID` and `KEGG` pathways/gene sets; (ii) a significant portion of miRNA–mRNA interactions have been experimentally validated in the literature; and (iii) identified miRNAs of each cancer type are significantly associated in the underlying cancer.

2 Materials

In this study, to find miRNA–mRNA regulatory modules, we used several cancer cohorts from the Cancer Genome Atlas (TCGA) project. The datasets used in our experiments are extracted from publicly available GDC data portal at https://portal.gdc.cancer.gov. TCGA project reported genomic characterizations and clinical information of 33 different cancer cohorts over more than 10 000 cancer patients.

Both miRNA and mRNA expression profiles are available in all cohorts except gliblastoma multiforme. Thus, we performed computational experiments on 32 cancer cohorts. We downloaded miRNA, mRNA and clinical data files from the most recent data freeze (i.e. `Data Release 26`—September `8, 2020`). We did not include metastatic tumors in our computational analyses since their underlying biology are different from primary tumors.

We extracted `BCGSC miRNA Profiling` files for miRNA expression profiles of all primary tumors which are preprocessed using the unified miRNASeq pipeline of TCGA. Total number of downloaded `BCGSC miRNA Profiling` files is 9975. We extracted `HTSeq-FPKM` files for mRNA expression profiles of all primary tumors which are preprocessed using the unified RNASeq pipeline of TCGA. Total number of downloaded `HTSeq-FPKM` files is 9755.

For our computational experiments, we needed the samples with matched miRNA and mRNA expression profiles, accordingly we eliminated primary tumors that have only one available expression profile. After this elimination step, total number of samples included in 32 cohorts is 9541. We presented a summary of the datasets used in our experiments for all 32 cohorts in Table 1. miRNA and mRNA profiles of tumors included expression values for 1881

Table 1. The details of 32 datasets constructed from TCGA cohorts

Cohort	Disease name	D	T	N_x	N_y	N
ACC	Adrenocortical carcinoma	577	16 913	80	79	79
BLCA	Bladder urothelial carcinoma	405	17 356	417	414	405
BRCA	Breast invasive carcinoma	538	17 530	1096	1102	1072
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	598	17 400	307	304	304
CHOL	Cholangiocarcinoma	548	17 315	36	36	36
COAD	Colon adenocarcinoma	532	17 002	455	478	441
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma	575	16 996	47	48	47
ESCA	Esophageal carcinoma	559	18 242	186	161	161
HNSC	Head and neck squamous cell carcinoma	609	17 509	523	500	495
KICH	Kidney chromophobe	513	17 064	66	65	65
KIRC	Kidney renal clear cell carcinoma	468	17 618	544	538	512
KIRP	Kidney renal papillary cell carcinoma	540	17 244	291	288	288
LAML	Acute myeloid leukemia	412	17 341	188	151	151
LGG	Brain lower grade glioma	616	17 628	512	511	507
LIHC	Liver hepatocellular carcinoma	579	16 805	372	371	367
LUAD	Lung adenocarcinoma	597	17 534	519	533	507
LUSC	Lung squamous cell carcinoma	587	17 748	478	502	475
MESO	Mesothelioma	559	17 418	87	86	86
OV	Ovarian serous cystadenocarcinoma	645	17 848	491	374	371
PAAD	Pancreatic adenocarcinoma	543	17 675	178	177	177
PCPG	Pheochromocytoma and paraganglioma	603	17 197	179	178	178
PRAD	Prostate adenocarcinoma	476	17 505	498	498	491
READ	Rectum adenocarcinoma	549	17 061	161	166	160
SARC	Sarcoma	499	17 248	259	259	257
SKCM	Skin cutaneous melanoma	651	17 201	97	103	97
STAD	Stomach adenocarcinoma	550	18 117	446	375	372
TGCT	Testicular germ cell tumors	798	18 068	150	150	150
THCA	Thyroid carcinoma	612	17 294	506	502	501
THYM	Thymoma	764	17 458	124	119	119
UCEC	Uterine corpus endometrial carcinoma	608	17 272	545	551	534
UCS	Uterine carcinosarcoma	682	17 809	57	56	56
UVM	Uveal melanoma	632	16 476	80	80	80
			Total	9975	9755	9541

Note: D is the number of miRNAs. T is the number of mRNAs. N_x is the number of primary tumors with available miRNA expression profiles. N_y is the number of primary tumors with available mRNA expression profiles. N is the number of primary tumors with available miRNA and mRNA expression profiles.

Downloaded from https://academic.oup.com/bioinformatics/article/36/Supplement_2/i592/6055910 by guest on 19 February 2021

miRNAs and 19 814 mRNAs. However, these numbers reduced to around 600 miRNAs and 17 000 mRNAs on the average after our filtering procedure, where we discarded miRNAs and mRNAs that were expressed in <50% of the tumors within each cohort from the analysis. The details about the numbers of miRNAs and mRNAs (i.e. D and T) included in each cancer type are also presented in Table 1.

To evaluate the biological relevance of the identified miRNA–mRNA regulatory modules, we performed survival analysis in our experiments. Toward this end, we also downloaded `clinical supplement` files of all patients. We then extracted their survival characteristics (i.e. days to last follow-up for alive patients and days to death for dead patients).

To find the biological relevance of the results, we used gene set collections. These collections present information about groups of genes which are similar or dependent in terms of their functions. We extracted the `Hallmark, PID` and `KEGG` pathways/gene sets from the Molecular Signatures Database (MSigDB), which can be publicly accessed at http://software.broadinstitute.org/gsea/msigdb.

3 Methods

We studied the problem of finding cancer-specific miRNA–mRNA regulatory modules from matched miRNA and mRNA expression profiles of tumors. In this problem, we are given a training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$, which contains N tumors together with their matched miRNA and mRNA levels, where $\boldsymbol{x}_i \in \mathbb{R}^D$ and $\boldsymbol{y}_i \in \mathbb{R}^T$ denote the miRNA and mRNA expression profiles of tumor i, respectively. Each training sample includes T mRNAs and D miRNAs expression profiles which can be represented as $\mathbf{Y} \in \mathbb{R}^{N\times T}$ and $\mathbf{X} \in \mathbb{R}^{N\times D}$ matrices, respectively.

We formulated this problem as a linear multivariate regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W} + \boldsymbol{E}, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{D\times T}$ is the matrix of regression coefficients and $\mathbf{E} \in \mathbb{R}^{N\times T}$ is the matrix of error terms. We considered miRNA and mRNA expression profiles as predictors and responses of the model, respectively. We investigated multivariate regression tasks between miRNA and mRNA expression profiles under the assumption that the columns of $\mathbf{X}$ and $\mathbf{Y}$ are centered (i.e. zero mean columns) and normalized (i.e. unit standard deviation columns). We assumed that the error terms in $\mathbf{E}$ are i.i.d. Gaussian random variables with zero mean and σ^2 variance.

In this high-dimensional problem, we assumed that the relationship between predictors (miRNAs) and responses (mRNAs) is sparse since the total number of responses and predictors are much larger than N, but the number of important factors is typically smaller than N. To fit the model on such data, regularized or penalized methods are needed to perform dimensionality reduction, and feature extraction is often the most important aspect of the analysis. Accordingly, we proposed a two-step feature extraction framework to efficiently find the sparse cancer-specific miRNA–mRNA regulatory modules from matched miRNA and mRNA expression profiles of tumors. The overall flow of the developed framework is demonstrated in Figure 1.

We developed a two-step algorithm:

1. In the first step, to estimate the coefficient matrix ($\mathbf{W}$) of the multivariate regression model between miRNA and mRNA expression profiles, we applied ℓ_1/ℓ_∞ norm on the cost function that decomposes the regression model to multiple linear programming problems. Then, we constructed $\mathbf{W}$ by concatenating the solution of each linear program together (Fig. 1c).
2. In the second step, we decomposed $\mathbf{W}$ into the product of two low-rank matrices ($\mathbf{W}_\mathcal{X}$ and $\mathbf{W}_\mathcal{Y}$). $\mathbf{W}_\mathcal{X}$ transforms predictors (i.e. miRNA profiles) to a set of latent factors and $\mathbf{W}_\mathcal{Y}$ regresses the responses (i.e. miRNA profiles) using the factors (Fig. 1d).

To this aim, we developed a non-convex RFR model with an elastic net penalty on both $\mathbf{W}_\mathcal{X}$ and $\mathbf{W}_\mathcal{Y}$ matrices to simultaneously extract low-rank sparse modules and also estimate effective

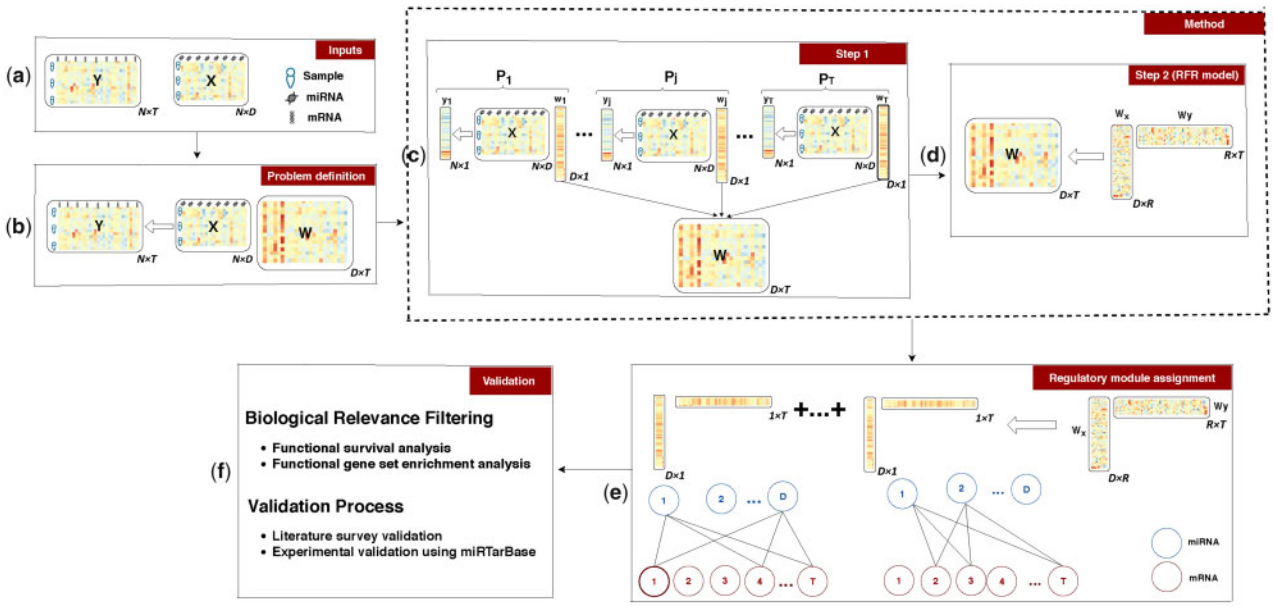


Fig. 1. Overview of the developed framework for identifying key miRNA–mRNA regulatory modules from their expression profiles. (**a**) Input data that include miRNA and mRNA expression profiles. (**b**) Linear multivariate regression model of miRNA–mRNA regulation. (**c**) Decomposing the addressed problem to T different linear modules and constructing coefficient matrix ($\mathbf{W}$) by concatenating them. (**d**) Imposing a low-rank structure by decomposing the obtained $\mathbf{W}$ matrix into multiplication of two low-rank matrices with effective estimation of R. (**e**) Writing the low-rank matrices as the summation of R rank-one matrices such that each of them corresponds to one miRNA–mRNA regulatory module. (**f**) Filtering procedure including functional gene set enrichment and survival analysis to filter the identified regulatory modules to key regulatory programs with biological relevance, as well as, validation process including literature survey to find the supporting evidences for miRNAs in the identified regulatory modules and experimental validation of identified interactions using miRTarBase dataset

Downloaded from https://academic.oup.com/bioinformatics/article/36/Supplement_2/i592/6055910 by guest on 19 February 2021

number of R. In addition to imposing sparse structure on $\mathbf{W}_\mathcal{X}$ and $\mathbf{W}_\mathcal{Y}$, the elastic net penalty groups correlated rows of $\mathbf{W}_\mathcal{X}$ together and columns of $\mathbf{W}_\mathcal{Y}$ together. To solve the proposed model, we provided a heuristic algorithm that decides the number of effective latent factors and efficiently estimates $\mathbf{W}_\mathcal{X}$ and $\mathbf{W}_\mathcal{Y}$ matrices. This low-rank sparse structure improves the feature extraction and interpretability of miRNA–mRNA regulatory modules. In the last part, to extract the identified miRNA–mRNA regulatory modules from low-rank matrices and sort them by their importance, we wrote the low-rank matrices as the summation of R rank-one matrices (Fig. 1e).

3.1 Step 1: ℓ_1/ℓ_∞ norm minimization

To find a structured relationship between miRNA and mRNA expression profiles, we developed a multiple linear programming formulation that effectively estimates the coefficient matrix of miRNA–mRNA regression model. Since regulation of a single mRNA by all miRNAs can be assumed as an independent task and there are T mRNA expression profiles, we decomposed these tasks into T independent tasks and solved them as separate problems.

For this purpose, to find the coefficient matrix of the regression problem (1), we considered ℓ_1/ℓ_∞ norm on the loss function of the miRNA–mRNA regression model as follows:

$$\begin{aligned} &\text{minimize} && \|\mathbf{Y}-\mathbf{X}\mathbf{W}\|_{1,\infty} \\ &\text{with respect to} && \mathbf{W}\in\mathbb{R}^{D\times T} \end{aligned} . \tag{2}$$

Each column of the response matrix $\mathbf{Y}$ indicates one single mRNA expression profile of all samples. By applying ℓ_1/ℓ_∞ norm definition, model (2) is equivalent to T independent problems as follows:

$$\begin{aligned} &\text{minimize} && \max_j \ |y_j-\boldsymbol{X}w_j| \\ &\text{with respect to} && w_j\in\mathbb{R}^D, \end{aligned} \tag{3}$$

where $j\in\{1,\ldots,T\}$, decision variables vector $\boldsymbol{w}_j\in\mathbb{R}^D$ denotes the jth column of the coefficient matrix ($\mathbf{W}$), and vector $\boldsymbol{y}_j\in\mathbb{R}^N$ indicates the jth column of known mRNA profiles ($\mathbf{Y}$).

Problem (3) can be expressed as linear programs. We composed the linear program P_j for jth mRNA (i.e. the jth column of $\mathbf{Y}$) as follows:

$$\begin{aligned} &\text{minimize} && t_j \\ &\text{with respect to} && t_j\in\mathbb{R},\ \ w_j\in\mathbb{R}^D \\ &\text{subject to} && t_j\mathbf{1}\geq y_j-\boldsymbol{X}w_j \\ & && -t_j\mathbf{1}\leq y_j-\boldsymbol{X}w_j \\ & && t_j\geq 0 \end{aligned} \tag{4}$$

where $t_j=\|\ \mathbf{y}_j-\mathbf{X}\mathbf{w}_j\ \|_\infty$ is the decision variable of the jth problem that is used to linearize the maximum of absolute values. We developed Problem (4) to estimate the jth column of the coefficient matrix ($\boldsymbol{w}_j$) of the miRNA–mRNA regression model using miRNA expression profiles of all N samples ($\mathbf{X}$) and N available expression profiles of the jth mRNA ($\boldsymbol{y}_j$) in a sparse structure. We constructed the coefficient matrix ($\mathbf{W}$) by concatenating all obtained $\boldsymbol{w}_j$ vectors after solving P_j for all $j\in\{1,\ldots,T\}$. Commercial solvers (e.g. Cplex, Gurobi and Mosek) can solve Problem (4) extremely fast even for a high-dimensional setting. This feature enabled us to find the optimal coefficient matrix by solving T linear programming problems.

3.2 Step 2: low-rank decomposition model

To select miRNA–mRNA regulatory modules, we formulated a low-dimensional structure that decomposes the obtained coefficient matrix of miRNA–mRNA regression model ($\mathbf{W}$) from Step 1, into two low-rank matrices with a sparse formulation and capability of capturing shared information among different miRNAs and mRNAs, and estimates the effective number of latent factors (R). To this aim, we designed an RFR model with element-wise sparsity term in the loss function and elastic net penalty as the regularization function. Applying the elastic net penalty, in addition to imposing a sparse structure, enabled our model to group correlated miRNAs together and correlated mRNAs together. Next, we proposed an alternating heuristic optimization algorithm to solve the resulting regularized model and find the number of effective latent factors.

It is a credible assumption that relationship between miRNAs and mRNAs is sparse, thus, to induce low-dimensionality structure, we decomposed $\mathbf{W}$ into multiplication of two low-rank matrices such that

$$\boldsymbol{W}=\boldsymbol{W}_\mathcal{X}\boldsymbol{W}_\mathcal{Y}+\boldsymbol{E}, \tag{5}$$

where $\mathbf{W}_\mathcal{X}\in\mathbb{R}^{D\times R}$ projects D-dimensional miRNA profiles into a R-dimensional space, $\mathbf{W}_\mathcal{Y}\in\mathbb{R}^{R\times T}$ performs linear regression in this R-dimensional projected space and $\mathbf{E}\in\mathbb{R}^{D\times T}$ is the matrix of error terms. Note that R is chosen a-priori in optimization such that $1\leq R\leq R_U\leq\min(D,T)$, where parameter R_U is the problem-specific upper bound for the dimensionality of the projected space.

This low-rank assumption, i.e. Problem (5), can be converted into a summation of rank-one matrices. We used each of these rank-one matrices as one miRNA–mRNA regulatory module in the further steps of our algorithm. For example, the first column of $\mathbf{W}_\mathcal{X}$ and the first row of $\mathbf{W}_\mathcal{Y}$ corresponded to the first miRNA–mRNA regulatory module.

To achieve the low-rank structure, we estimated $\mathbf{W}_\mathcal{X}$ and $\mathbf{W}_\mathcal{Y}$ from $\mathbf{W}$ by a regularized model that minimizes least absolute error of the model (5) as follows:

$$\begin{aligned} &\text{minimize} && \|\boldsymbol{W}-\boldsymbol{W}_\mathcal{X}\boldsymbol{W}_\mathcal{Y}\|_{1,1}+\mathcal{L}_\mathcal{X}(\boldsymbol{W}_\mathcal{X})+\mathcal{L}_\mathcal{Y}(\boldsymbol{W}_\mathcal{Y}) \\ &\text{with respect to} && \boldsymbol{W}_\mathcal{X}\in\mathbb{R}^{D\times R},\ \ \boldsymbol{W}_\mathcal{Y}\in\mathbb{R}^{R\times T}, \end{aligned}$$

where $\mathcal{L}_\mathcal{X}(\cdot)$ and $\mathcal{L}_\mathcal{Y}(\cdot)$ are the regularization functions defined on $\mathbf{W}_\mathcal{X}$ and $\mathbf{W}_\mathcal{Y}$, respectively, to restrict the search space and to enforce a sparse structure on the variables.

The Lasso is a popular regularization technique for simultaneous estimation and variable selection and, same as other ℓ_1 regularized methods, is computationally feasible in high-dimensional problems (Efron *et al.*, 2004). However, there exist definite scenarios that Lasso is inconsistent for variable selection. Zou (2006) derived a necessary condition for the Lasso variable selection to be consistent by proposing a model called adaptive Lasso that minimizes the regularized mean square error with a Frobenius norm on $\mathbf{W}_\mathcal{X}$ and an orthonormality constraint on $\mathbf{W}_\mathcal{Y}$ (i.e. $\mathbf{W}_\mathcal{Y}\mathbf{W}_\mathcal{Y}^T=\mathbf{I}$), which leads to uncorrelated feature extraction.

To improve the interpretability of the feature extraction, Zou *et al.* (2006) introduced the sparse principal component analysis (SPCA) method by adding the ℓ_1-norm penalty on $\mathbf{W}_\mathcal{X}$ of the adaptive Lasso. However, adding the ℓ_1-norm penalty on $\mathbf{W}_\mathcal{X}$ makes the identified modules correlated. We formulated the regularization functions of our model inspired by the SPCA method with two main differences:

1. We imposed ℓ_1 and Frobenius norm penalties (i.e. elastic net penalty) on both $\mathbf{W}_\mathcal{X}$ and $\mathbf{W}_\mathcal{Y}$, which offers higher level of sparsity and interpretability that can be controlled by changing the regularization parameters.
2. We also discarded orthonormality constraint and added full rank condition to our problem to make sure that R columns of $\mathbf{W}_\mathcal{X}$ and R rows of $\mathbf{W}_\mathcal{Y}$ are linearly independent, which leads to selecting uncorrelated modules.

Based on a theorem developed by Zou and Hastie (2005), the elastic net penalty enforces the correlated features to group together. In other words, we expect two highly correlated features to both exist or both absent in a factor. To guarantee identifying linearly independent modules, we used the mechanism proposed by Kharratzadeh and Coates (2016) to learn from data that how many independent latent factors are needed to explain the data. This helped us to avoid both overfitting and underfitting, which could occur due to choosing a large or small number of hidden factors (R), respectively.

Downloaded from https://academic.oup.com/bioinformatics/article/36/Supplement_2/i592/6055910 by guest on 19 February 2021

Accordingly, by considering the elastic net penalty, which linearly combines $\mathcal{L}_1$ and Frobenius norms, as $\mathcal{L}_{\mathcal{X}}(\cdot)$ and $\mathcal{L}_{\mathcal{Y}}(\cdot)$, we obtained the regularized decomposition model as

$$\begin{aligned}\text{minimize} \quad & \|\mathbf{W}-\mathbf{W}_{\mathcal{X}}\mathbf{W}_{\mathcal{Y}}\|_{1,1}+\lambda_1\|\mathbf{W}_{\mathcal{X}}\|_{1,1}+\lambda_2\|\mathbf{W}_{\mathcal{X}}\|_F^2\,\mathcal{L}_{\mathcal{X}}(\mathbf{W}_{\mathcal{X}}) \\ & +\lambda_3\|\mathbf{W}_{\mathcal{Y}}\|_{1,1}+\lambda_4\|\mathbf{W}_{\mathcal{Y}}\|_F^2\,\mathcal{L}_{\mathcal{Y}}(\mathbf{W}_{\mathcal{Y}}) \text{with respect to } \mathbf{W}_{\mathcal{X}} \\ & \in\mathbb{R}^{D\times R},\ \mathbf{W}_{\mathcal{Y}}\in\mathbb{R}^{R\times T}\end{aligned} \tag{6}$$

where $\lambda_1\in\mathbb{R}_+$ and $\lambda_2\in\mathbb{R}_+$ are the user-defined regularization parameters of the elastic net penalty on $\mathbf{W}_{\mathcal{X}}$, and, similarly, $\lambda_3\in\mathbb{R}_+$ and $\lambda_4\in\mathbb{R}_+$ are regularization parameters on $\mathbf{W}_{\mathcal{Y}}$. Problem (6) is non-convex, hence, it is difficult to solve it exactly in a reasonable time. However, with a predetermined R, quadratic penalty term on both $\mathbf{W}_{\mathcal{X}}$ and $\mathbf{W}_{\mathcal{Y}}$ makes Problem (6) strongly convex, if either $\mathbf{W}_{\mathcal{X}}$ or $\mathbf{W}_{\mathcal{Y}}$ is fixed. This feature enabled us to propose a heuristic optimization algorithm to efficiently estimate the solution of Problem (6). Our proposed algorithm starts with an initial upper bound on the rank (i.e. $R=R_U$) and solves the optimization problem (6) using an alternative strategy since the objective function is not jointly convex with respect to $\mathbf{W}_{\mathcal{X}}$ and $\mathbf{W}_{\mathcal{Y}}$, but it is convex with respect to $\mathbf{W}_{\mathcal{X}}$ if we fix $\mathbf{W}_{\mathcal{Y}}$ or vice versa. Thus, after fixing one of the matrices, Problem (6) can be expressed as a quadratic programming problem, which can also be solved for high-dimensional problems. After convergence, the algorithm checks whether both $\mathbf{W}_{\mathcal{X}}$ and $\mathbf{W}_{\mathcal{Y}}$ are full rank. If both of them are not full rank, it reduces the value of the R parameter by one and solves the optimization problem (6) again. At the end of this iterative procedure, the algorithm is guaranteed to have full rank $\mathbf{W}_{\mathcal{X}}$ and $\mathbf{W}_{\mathcal{Y}}$ matrices. This optimization strategy is described in the Supplementary Material with more details.

3.3 miRNA–mRNA regulatory module assignment

After finding $\mathbf{W}_{\mathcal{X}}$ and $\mathbf{W}_{\mathcal{Y}}$ by running the algorithm, we specified weights to each identified regulatory module. Our basic assumption for decomposition was $\mathbf{W}\approx\mathbf{W}_{\mathcal{X}}\mathbf{W}_{\mathcal{Y}}$. To have a unified scale for all of the modules, we normalized each row of $\mathbf{W}_{\mathcal{X}}$ and each column of $\mathbf{W}_{\mathcal{Y}}$ to unit norm. We obtained the low-rank decomposition $\mathbf{W}\approx\tilde{\mathbf{W}}_{\mathcal{X}}\mathbf{D}_{\mathcal{X}}\mathbf{D}_{\mathcal{Y}}\tilde{\mathbf{W}}_{\mathcal{Y}}$, where $\mathbf{D}_{\mathcal{X}}$ and $\mathbf{D}_{\mathcal{Y}}$ are $R\times R$ diagonal matrices. The diagonal entries of the multiplication of $\mathbf{D}_{\mathcal{X}}$ and $\mathbf{D}_{\mathcal{Y}}$ are used to assign respective importance weights to the regulatory modules. We then sorted the regulatory modules by sorting the columns of $\mathbf{W}_{\mathcal{X}}$ and the rows of $\mathbf{W}_{\mathcal{Y}}$ from the largest to the smallest. The regulatory module that corresponds to the highest diagonal entry is the regulatory module with the highest importance.

To check whether our algorithm finds meaningful miRNA–mRNA regulatory modules, we determined selected miRNAs and mRNAs for each module. We thresholded the columns of $\tilde{\mathbf{W}}_{\mathcal{X}}$ and the rows of $\tilde{\mathbf{W}}_{\mathcal{Y}}$ as follows: an miRNA is considered to be selected if the magnitude of its weight is larger than one over square root of the total number of miRNAs included, and an mRNA is considered to be selected if the magnitude of its weight is larger than one over square root of the total number of mRNAs included. We then clustered the patients into two groups using k-means clustering on the expression values of the selected mRNAs only. To select key regulatory modules among all identified miRNA–mRNA regulatory modules, we filtered identified modules by two different functional analysis (i.e. enrichment analysis and survival analysis).

First, to evaluate the biological relevance of identified regulatory modules, we performed a functional gene set enrichment analysis. To this aim, we found the enrichment of identified regulatory modules in the `Hallmark`, `PID` and `KEGG` collections using the hypergeometric test.

Then, we investigated the functional importance of identified miRNA–mRNA regulatory modules by checking whether there is a statistically significant survival difference between the two groups of patients found by the clustering algorithm using the log-rank test. A regulatory module is considered to be selected as a key regulatory module if the following conditions are met.

- The identified regulatory module significantly enriched in at least one of the pathway/gene set collections.
- Survival analysis using the identified module reports a statistically significant difference between patients' groups.

4 Results and discussion

4.1 Experimental setting

To determine reasonable values for the regularization parameters of Problem (6), i.e. λ_1, λ_2, λ_3 and λ_4, we randomly picked 50 miRNAs and 200 mRNAs of each cohort. We considered seven possible values for each parameter (i.e. 2, 3, 4, 5, 6, 7 and 8) and calculated the root mean square error (RMSE) for each parameter set on all cancer cohorts. Finally, a set of four parameters with minimum average error on all datasets has been selected as the regularization parameters of the full-scale experiments. The selected values of λ_1, λ_2, λ_3 and λ_4 were 6, 2, 8 and 3, respectively. We considered $R_U=20$ as the upper bound on the number of latent factors.

We determined the stopping criterion of the algorithm based on the objective function of the optimization problem (6). We considered the ratio $|f^{(t+1)}-f^{(t)}|/f^{(t)}$ as the convergence criterion, where $f^{(t+1)}$ and $f^{(t)}$ are the objective function values of the last two iterations. The algorithm terminates when $|f^{(t+1)}-f^{(t)}|/f^{(t)}<\epsilon$, where $\epsilon=10^{-5}$. We implemented our algorithms in R, which uses `MOSEK` version 9.1.5 to solve quadratic optimization problems (MOSEK ApS, 2019).

4.2 Predictive performance comparison

As discussed, the most notable statistical methods that are introduced for identifying cancer-specific miRNA regulatory modules rely on NMF or SVD. Note that since the relationships between miRNAs and mRNAs are not modeled in the NMF-based methods, we cannot compare their predictive performance against our proposed two-step framework. However, we can compare our method with SVD-based approaches in terms of NRMSE. We compared the performance of our algorithm with T-SVD regression (Ma *et al.*, 2014), which has an available R package and has also shown better performance against similar algorithms. Note that T-SVD outperformed the sequential extraction algorithm and iterative exclusive extraction algorithm introduced earlier by Chen *et al.* (2012), and two biclustering-based SVD methods (Lee *et al.*, 2010; Yang *et al.*, 2014) in terms of sum squared error.

We compared the predictive performance of our method against T-SVD regression in terms of their NRMSE between observed and predicted values of mRNA expression level calculated using miRNA expression levels and weights matrices (i.e. $\hat{\mathbf{Y}}=\mathbf{X}\mathbf{W}_{\mathcal{X}}\mathbf{W}_{\mathcal{Y}}$). To this aim, we applied T-SVD regression on the same data using `tsvd` R package (Ma *et al.*, 2015).

The average NRMSE values of our algorithm and T-SVD regression over mRNAs of all cancer types are reported in Table 2. The predictive performance values showed that our two-step algorithm was able to explain a higher proportion of variance compared to T-SVD regression on 29 out of 32 cohorts. Our algorithm and T-SVD regression were not able to learn in PRAD and THYM cohorts, respectively.

4.3 Extracting key miRNA–mRNA regulatory modules

We applied k-means clustering on the expression values of the selected mRNAs to cluster the patients into two groups. We then obtained the expression values of selected miRNAs and mRNAs together with the clustering assignment of patients. Figures 2 and 3 show two example heat maps of selected miRNAs and mRNAs for LIHC and PAAD cohorts, respectively. In these heat maps, red colors show over-expression (i.e. higher than the sample mean) and blue colors show under-expression (i.e. lower than the sample

Downloaded from https://academic.oup.com/bioinformatics/article/36/Supplement_2/i592/6055910 by guest on 19 February 2021

Table 2. Predictive performance values of our algorithm and T-SVD regression algorithm on 32 datasets in terms of average NRMSE over mRNAs and their selected ranks

	T-SVD		Our algorithm	
Cohort	Rank	NRMSE	Rank	NRMSE
ACC	5	0.9592	18	**0.7188**
BLCA	21	0.9321	20	**0.8666**
BRCA	44	**0.9230**	20	0.9419
CESC	19	0.9327	20	**0.8470**
CHOL	2	0.9842	16	**0.5304**
COAD	23	0.9321	20	**0.8151**
DLBC	2	0.9868	19	**0.5803**
ESCA	9	0.9864	20	**0.7995**
HNSC	27	0.9198	20	**0.8549**
KICH	4	0.9871	18	**0.5906**
KIRC	27	**0.9187**	20	0.9796
KIRP	17	0.9081	20	**0.7838**
LAML	10	0.9845	20	**0.7479**
LGG	20	0.9043	20	**0.8267**
LIHC	22	0.9106	20	**0.8197**
LUAD	27	0.9304	20	**0.8832**
LUSC	28	0.9759	20	**0.8870**
MESO	6	0.9929	20	**0.7180**
OV	22	0.9212	20	**0.8602**
PAAD	10	0.9940	20	**0.7227**
PCPG	12	0.9359	20	**0.7577**
PRAD	22	**0.9202**	20	1.0016
READ	8	0.9999	20	**0.7304**
SARC	14	0.8962	20	**0.8788**
SKCM	7	0.9513	20	**0.6851**
STAD	21	0.9317	20	**0.8378**
TGCT	8	0.8513	13	**0.7040**
THCA	23	0.9142	20	**0.7983**
THYM	6	1.0226	13	**0.6513**
UCEC	26	0.9359	20	**0.8567**
UCS	3	0.9813	18	**0.6664**
UVM	5	0.9308	19	**0.5286**

Note: Better performance values are shown with bold font.

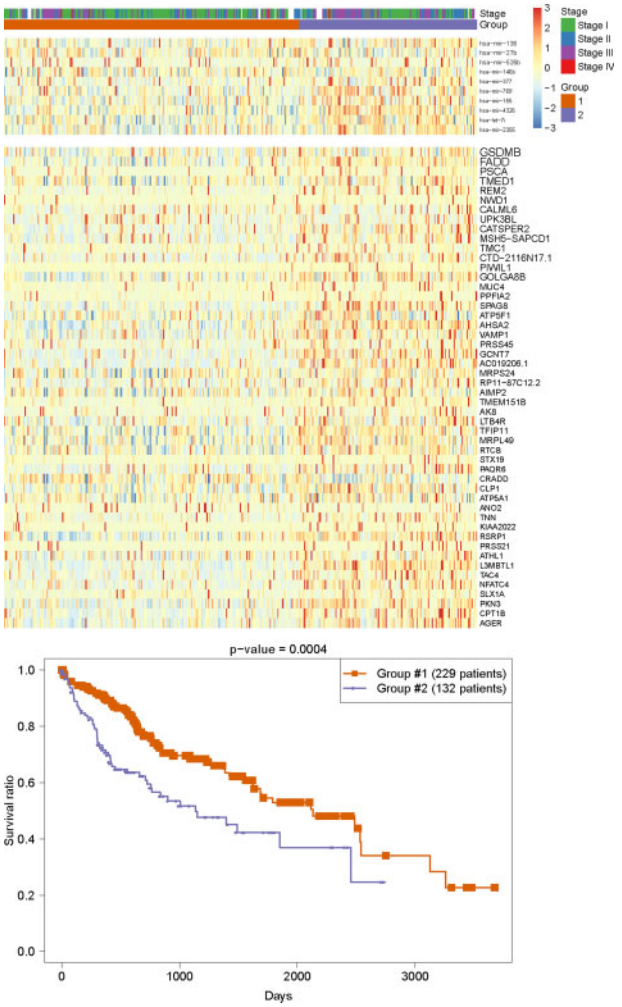


Fig. 2. Regulatory module no. 8 identified for LIHC cohort. The heat map shows the expression profiles of top 10 miRNAs and top 50 mRNAs. The Kaplan–Meier curves show survival characteristics of two patient groups

mean). The font sizes for the names of miRNAs and mRNAs are proportional to the magnitudes of their weights inferred by our algorithm. For cohorts with available pathological stage information, the pathological stage of each patient has also been reported in the heat maps. However, we have not observed any clear connection between patient group and pathological stage.

To extract key regulatory modules and to check whether our algorithm finds valid miRNA–mRNA regulatory modules, we filtered the identified regulatory modules by enrichment and survival analyses. In the first stage of the filtering process, we performed functional gene set enrichment analysis to see whether the identified modules are significantly enriched in known functional `Hallmark`, `PID` and `KEGG` collections. We calculated enrichment of 614 identified miRNA–mRNA regulatory modules in total, in the `Hallmark`, `PID` and `KEGG` collections using the hypergeometric test. We found at least one enriched gene set in the `Hallmark`, `PID` or `KEGG` collections with an FDR-corrected q-value <0.05 in 432 out of 614 (70.36%) regulatory modules.

Next, we examined clinical characterization of identified miRNA–mRNA regulatory modules by checking whether there is a significant survival difference between the two groups of patients found by the clustering algorithm using the log-rank test. We checked 614 miRNA–mRNA regulatory modules, and, in only 146 of them, we observed a significant survival difference between the two groups (i.e. P-value <0.05 in the log-rank test). Figures 2 and 3 show two example Kaplan–Meier survival curves for LIHC and PAAD cohorts, respectively. Consequently, 97 out of 614 modules were selected by both survival and enrichment analyses as key regulatory modules. These results show the power of our method in grouping genes that participate in the same biological processes or gene sets.

4.4 Literature validation of identified miRNA–mRNA regulatory modules

To validate the results of our experiment, we checked the existing literature to find cancer-specific supporting evidence for 97 key miRNA–mRNA regulatory modules identified by our algorithm. We searched the PubMed repository at https://www.ncbi.nlm.nih.gov/pubmed using top ten miRNA names in terms of the magnitude of their weights and cancer name. We were able to find supporting evidence for 282 miRNAs from 79 out of 97 (81.44%) regulatory modules of underlying cancer type. The publications that provide supporting evidence for miRNA–mRNA regulatory modules identified by our algorithm are listed in Supplementary Table S1. For 79 regulatory modules with significant survival difference and gene set enrichment, detailed results of gene set enrichment analysis have been reported in Supplementary Table S2. We also reported the miRNA and mRNA expression heat maps and Kaplan–Meier survival curves in the Supplementary Material.

We also checked the matched miRNA–mRNA interactions identified by our model and T-SVD regression with experimentally validated miRNA–mRNA interactions in miRTarBase dataset (Chou *et al.*, 2018). We picked top five miRNAs and top 50 mRNAs in

Downloaded from https://academic.oup.com/bioinformatics/article/36/Supplement_2/i592/6055910 by guest on 19 February 2021

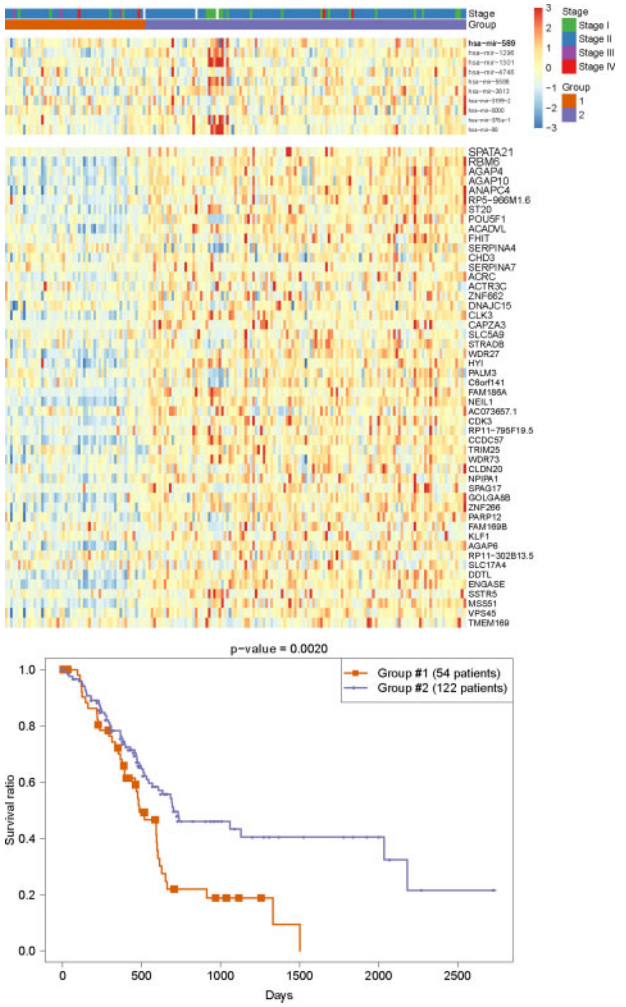


Fig. 3. Regulatory module no. 8 identified for PAAD cohort. The heat map shows the expression profiles of top 10 miRNAs and top 50 mRNAs. The Kaplan–Meier curves show survival characteristics of two patient groups

Table 3. Numbers and percentages of interactions identified by our algorithm and T-SVD regression algorithm validated by miRTarBase dataset on 32 datasets

	T-SVD		Our algorithm	
Cohort	Number	Percentage	Number	Percentage
ACC	0	0.0000	1	**0.0222**
BLCA	1	0.0513	14	**0.2800**
BRCA	11	**0.2340**	0	0.0000
CESC	2	0.1053	17	**0.3400**
CHOL	4	**0.2963**	5	0.1250
COAD	3	**0.1429**	0	0.0000
DLBC	0	0.0000	13	**0.2737**
ESCA	0	0.0000	10	**0.2000**
HNSC	3	**0.1224**	2	0.0400
KICH	0	0.0000	14	**0.3111**
KIRC	3	**0.1132**	0	0.0000
KIRP	3	**0.1622**	6	0.1200
LAML	2	**0.2000**	1	0.0200
LGG	2	0.0980	5	**0.1000**
LIHC	3	**0.1580**	5	0.1000
LUAD	7	**0.3180**	2	0.0000
LUSC	4	**0.1702**	1	0.0200
MESO	0	0.0000	3	**0.0600**
OV	9	0.0865	7	**0.1400**
PAAD	0	0.0000	5	**0.1000**
PCPG	1	0.1176	14	**0.2800**
PRAD	4	**0.1455**	0	0.0000
READ	0	0.0000	5	**0.1000**
SARC	9	0.1208	9	**0.1800**
SKCM	2	**0.2500**	5	0.1000
STAD	4	**0.2000**	4	0.0800
TGCT	5	**0.1316**	3	0.0923
THCA	1	**0.0513**	2	0.0400
THYM	7	0.1770	7	**0.2154**
UCEC	0	0.0000	1	**0.0200**
UCS	2	0.1379	7	**0.1556**
UVM	0	0.0000	15	**0.3158**

Note: Better performance values are shown with bold font.

terms of the magnitude of their weights, identified by our algorithm and by T-SVD regression, and we extracted matched interactions of both methods with miRTarBase dataset interactions. We found 92 and 183 validated interactions by T-SVD regression and our method, respectively, that have been listed in Supplementary Tables S3 and S4. We compared the percentages of validated interactions for our algorithm and T-SVD regression in all 32 cohorts. Our algorithm outperformed T-SVD regression in terms of validated interaction percentage in 17 out of 32 cohorts. The detailed percentages of validated interactions of our algorithm and T-SVD regression have been reported in Table 3.

5 Conclusions

miRNAs play critical role in gene regulation of many biological processes such as formation and progression of cancer. Since only a limited number of miRNAs regulating to mRNAs are found experimentally, identifying miRNA–mRNA regulatory modules are crucial to the process of revealing their biological function.

In this article, we introduced a framework to efficiently identify and analyze miRNA–mRNA regulatory modules from their matched expression profiles. To extract the effective cancer-related miRNAs, we formulated the relationship between miRNAs and mRNAs in a two-step algorithm. We applied our proposed two-step algorithm on TCGA cancer data with matched miRNA and mRNA expression profiles of primary tumors for 32 out of 33 cancer types, which makes these experiments, to the best of our knowledge, the most comprehensive study on identifying miRNA–mRNA regulatory modules.

To demonstrate the effectiveness of our proposed method, we compared the predictive performance of our algorithm with T-SVD regression in terms of average NRMSE over mRNAs. Our results showed that our two-step algorithm was able to explain significantly larger portions of variance on 29 out of 32 datasets. To validate and show the biological relevance of the extracted information, we checked whether there is any supporting evidence from the literature for the modules that have a significant survival difference between patient groups. We then filtered out the identified modules by testing whether the identified modules are significantly enriched in known functional `Hallmark`, `PID` and `KEGG` pathway/gene set collections.

Our framework was able to find 97 key regulatory modules, and 79 of them have supporting cancer-related evidence from the literature (Supplementary Table S1). We reported heat maps of miRNA and mRNA expression values together with the Kaplan–Meier survival curves, and also functional gene set enrichment analysis results for these 79 regulatory modules (Figs 2 and 3; Supplementary Figs S1–S77 and Table S2). We also checked miRTarBase dataset to find experimentally validated evidences of the identified miRNA–mRNA interactions by our algorithm and compared against that of T-SVD regression. We reported identified miRNA–mRNA interactions matched with miRTarBase in Supplementary Tables S3 and S4.

We envision an extension of this study toward considering more genomic data sources, in addition to miRNA and mRNA expression

Downloaded from https://academic.oup.com/bioinformatics/article/36/Supplement_2/i592/6055910 by guest on 19 February 2021

profiles such as gene–gene interactions and miRNA–gene predicted interactions that have important roles in miRNA–mRNA regulatory modules. Our computational framework can be applied to integrate multiple data sources and extract key miRNA–mRNA regulatory modules with some minor modifications.

Acknowledgements

Computational experiments were performed on the OHSU Exacloud high performance computing cluster.

Funding

Mehmet Gönen was supported by the Turkish Academy of Sciences (TÜBA-GEBİP; The Young Scientist Award Program) and the Science Academy of Turkey (BAGEP; The Young Scientist Award Program).

Conflict of Interest: none declared.

References

Ajana,S. *et al.* (2019) Benefits of dimension reduction in penalized regression methods for high-dimensional grouped data: a case study in low sample size. *Bioinformatics*, **35**, 3628–3634.

Argyriou,A. *et al.* (2008) Convex multi-task feature learning. *Mach. Learn.*, **73**, 243–272.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Bonnet,E. *et al.* (2010) Module network inference from a cancer gene expression data set identifies microRNA regulated modules. *PLoS One*, **5**, e10162.

Breheny,P. (2015) The group exponential lasso for bi-level variable selection. *Biometrics*, **71**, 731–740.

Calin,G.A. and Croce,C.M. (2006) MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, **6**, 857–866.

Chen,K. *et al.* (2012) Reduced rank stochastic regression with a sparse singular value decomposition. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, **74**, 203–221.

Chen,L. and Huang,J.Z. (2012) Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Am. Stat. Assoc.*, **107**, 1533–1545.

Chou,C.H. *et al.* (2018) miRTarBase update 2018: a resource for experimentally validated microRNA–target interactions. *Nucleic Acids Res.*, **46**, D296–D302.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–499.

Golub, G.H. and Reinsch, C. (1971) Singular value decomposition and least squares solutions. In: Bauer,F.L. *et al.* (eds), *Handbook for Automatic Computation: Volume II: Linear Algebra*. Springer, Berlin, Heidelberg, pp. 134–151.

Hao,X. *et al.* (2017) Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis. *Bioinformatics*, **33**, i341–i349.

He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, **5**, 522–531.

Joung,J.-G. *et al.* (2007) Discovery of microRNA–mRNA modules via population-based probabilistic learning. *Bioinformatics*, **23**, 1141–1147.

Kharratzadeh,M. and Coates,M. (2016) Sparse multivariate factor regression. In: *IEEE Statistical Signal Processing Workshop,* Palma de Mallorca, Spain June 2016.

Koo,J. *et al.* (2018) Tiresias: context-sensitive approach to decipher the presence and strength of microRNA regulatory interactions. *Theranostics*, **8**, 277–291.

Kumar,A. and Daume,H.,III (2012) Learning task grouping and overlap in multi-task learning. In: *29th International Conference on Machine Learning,* Edinburgh, Scotland June 2012.

Lee,M. *et al.* (2010) Biclustering via sparse singular value decomposition. *Biometrics*, **66**, 1087–1095.

Li,Y. *et al.* (2014) A probabilistic approach to explore human miRNA targetome by integrating miRNA-overexpression data and sequence information. *Bioinformatics*, **30**, 621–628.

Ma,X. *et al.* (2014) Learning regulatory programs by threshold SVD regression. *Proc. Natl. Acad. Sci. USA*, **111**, 15675–15680.

Ma,X. *et al.* (2015) *tsvd: Thresholding-based SVD for Multivariate Reduced Rank Regression*. R Package Version 1.4, March 27, 2015.

Massy,W.F. (1965) Principal components regression in exploratory statistical research. *J. Am. Stat. Assoc.*, **60**, 234–256.

Min,W. *et al.* (2016) A two-stage method to identify joint modules from matched microRNA and mRNA expression data. *IEEE Trans. Nanobiosci.*, **15**, 362–370.

MOSEK ApS. (2019) *MOSEK Optimization Suite Release Version 9.1.5, October 29, 2019*.

Peng,J. *et al.* (2010) Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, **4**, 53–77.

Raad,J. *et al.* (2020) Complexity measures of the mature miRNA for improving pre-miRNAs prediction. *Bioinformatics*, **36**, 2319–2327.

Shi,Q. *et al.* (2017) Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. *Bioinformatics*, **33**, 2706–2714.

Wang,Z. *et al.* (2015) Integrating full spectrum of sequence features into predicting functional microRNA–mRNA interactions. *Bioinformatics*, **31**, 3529–3536.

Xiao,Q. *et al.* (2019) CeModule: an integrative framework for discovering regulatory patterns from genomic data in cancer. *BMC Bioinformatics*, **20**, 67.

Yang,D. *et al.* (2014) A sparse singular value decomposition method for high-dimensional data. *J. Comput. Graph. Stat.*, **23**, 923–942.

Zhang,C.-H. and Huang,J. (2008) The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Stat.*, **36**, 1567–1594.

Zhang,S. *et al.* (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA–gene regulatory modules. *Bioinformatics*, **27**, i401–i409.

Zhang,Y. *et al.* (2019) Genome-wide identification of the essential protein-coding genes and long non-coding RNAs for human pan-cancer. *Bioinformatics*, **35**, 4344–4349.

Zou,H. (2006) The adaptive Lasso and its oracle properties. *J. Am. Stat. Assoc.*, **101**, 1418–1429.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 301–320.

Zou,H. *et al.* (2006) Sparse principal component analysis. *J. Comput. Graph. Stat.*, **15**, 265–286.

Downloaded from https://academic.oup.com/bioinformatics/article/36/Supplement_2/i592/6055910 by guest on 19 February 2021